Experimental Plans in Factorial Surveys: Random or Quota Design?

Dülmer, Hermann

Veröffentlichungsversion / Published Version
Zeitschriftenartikel / journal article

Sociological Methods
& Research
Volume 35 Number 3
February 2007 382-409
© 2007 Sage Publications
10.1177/0049124106292367
http://smr.sagepub.com
hosted at
http://online.sagepub.com

Experimental Plans in Factorial Surveys

Random or Quota Design?

Hermann Dülmer
University of Cologne, Germany

The factorial survey is an experimental design where respondents are asked to judge descriptions of varying situations (vignettes) presented to them. Combining the vignette variables (factors) and their levels is done by the researcher, who also takes the responsibility for getting an optimal design. To represent the universe of possible level combinations as accurately as possible, random designs are mostly used. A possible alternative are quoted designs. Up to now, there has been little discussion and only few research studies done about the pros and cons of random and quota samples for factorial surveys. The purpose of this article is to contribute to filling this gap. The conclusions drawn from the statistical considerations are illustrated by example analyses on the basis of fictitious data. Since the data structure produced by a factorial survey is a hierarchical one, the empirical analyses are carried out by using a multilevel program.

Keywords: *factorial survey; vignettes; random design; fractional factorial design; D-efficient design; multilevel analysis*

The factorial survey—also known as vignette analysis—is a procedure that allows one to analyze judgment behavior under concrete conditions that are much closer to real-life judgment-making situations than relatively abstract questions that are more typical for opinion surveys. Judging several of such similar but not identical situations by each respondent allows decomposing the structure of the individual answer behavior and thereby uncovering the impact of its different determinants. The central idea behind factorial surveys, a procedure that essentially goes back to Rossi (cf. Rossi and Nock 1982:9 or Jasso 1988:921), consists of transferring the basic principles of the factorial design into the sample survey.

Author's Note: I thank two anonymous *SMR* reviewers for a number of helpful comments and suggestions.

The factorial design is an experimental design in which the researcher constructs some descriptions of similar situations, which will be judged by respondents under a particular aspect. One could, for instance, examine on the basis of concrete situational descriptions how much age, education, knowledge of a country's language, nationality/origin, and religion of a fictitious immigrant influence the extent respondents agree or disagree with providing a residence permit or nationalization into one's own country (for a similar research question, see Jasso 1988). A respective situational description (vignette) can be presented either on a record sheet or a rectangular field in the questionnaire. A vignette for judging the social status of a person could, for example, include one of three specified levels for education as well as three levels for occupation (cf. Rossi and Anderson 1982:31). For two variables, each consisting of three different levels, it is possible to construct a maximum number of $3^2 = 9$ vignettes (Cartesian product), which together represent the completely crossed vignette universe. Judging all nine vignettes by each participant ensures that the given variables stand according to their composition orthogonal to each other. The designation *factorial* design also goes back to this feature.

With an increasing number of characteristics considered important from the theoretical point of view and/or with an increasing number of levels distinguished for each characteristic, it rapidly gets impossible for a respondent to judge all vignettes of the complete vignette universe. To bridge this bottleneck, the demand for judging all vignette combinations by each participant has been relaxed for the *factorial survey*. In factorial surveys, each respondent has to judge only a reduced sample from the entire vignette universe. The number of vignettes included into the reduced sample should remain high enough for estimating respondent-specific regression analyses. In this way, one of the central advantages of factorial surveys, the possibility for decomposing the structure of individual answer behavior, will be retained. However, by reducing the sample size, the question of getting an optimal sample becomes important.

Basically, two different designs for arranging vignette samples have been suggested in the literature: These are *random designs* and *quota designs*. Most introductions into vignette analysis only advise drawing random samples (e.g., Rossi 1979:179; Rossi and Anderson 1982:40-41; Jasso 2006:343), whereby each participant gets a unique random sample of the same size. A basically different procedure for constructing the samples is predominantly used for conjoint analysis, a method very closely related to vignette analysis:[1] Conjoint analyses are carried out almost always with special *quota samples*, among them predominantly the so-called fractional

factorial designs (cf. Gustafsson, Herrmann, and Huber 2000:17; Marshall and Bradlow 2002:675). Alternatively, one also can use another kind of quota sample, the so-called D-efficient designs (cf. Kuhfeld, Tobias, and Garratt 1994). In both of these cases, all participants get identical sets of vignettes to be judged.

Up to now, there has been little discussion and only a few research studies about the pros and cons of random and quota samples for factorial surveys.[2] The purpose of this contribution is—besides giving a short introduction to the basic ideas of their main variants—to compare both methods regarding their main advantages and disadvantages with respect to feasibility as well as to statistical and methodological issues. The conclusions drawn from the latter considerations will be illustrated by example multilevel analyses carried out by using simulated answer behavior for random as well as for quota designs.[3]

Basic Ideas and Main Variants of Random and Quota Designs

Random Designs

First, two proposals that have been made in the literature for using *random designs* can be distinguished. One proposal recommended by Rossi and Anderson (1982:40-1), in their introduction to the factorial survey, is to pick out the values (levels) of each vignette variable (characteristic) at random. In this way, a unique random sample of the same extent will be produced for each participant (*simple random design with replacement*; for an application, see, e.g., Alves and Rossi 1978:544-5). To avoid having a vignette selected twice for a vignette set, it might be preferable in general to draw the vignettes for each vignette set separately at random from the fully crossed vignette universe (*simple random design without replacement*; cf. also Jasso 2006:342-3). The basic idea behind using such simple random designs is to represent the complete vignette universe as accurately as possible by distinct vignette samples of the same sample size. Within the limits of sampling error, each randomly drawn vignette sample is a reduced representative sample of the whole vignette sample from which it originates. Adding such random samples to each other produces again a random sample of the whole vignette universe. Thus, the combined random sample is like the individual distinct random samples within the limits of a comparably much smaller sampling error, which is

also a representative sample of the whole vignette universe (cf. Rossi and Anderson 1982:29-30).

The other proposal, more recently made by Beck and Opp (2001:292-3), is to draw without replacement only several random samples of the same size from the whole vignette universe and to use each of these vignette sets several times for an advance specified, fixed number of different participants (*clustered random design*). Probably, the most important reason for applying this variant is to obtain multiple ratings per vignette, allowing not only respondent-specific but also vignette-specific analyses (cf. also Jasso 2006:379-80). Drawing vignettes without replacement ensures that a maximum number of different vignettes will be covered by the combined vignette sample. If the completely crossed vignette universe consists of relatively few different vignettes, if the chosen set size is relatively high, and/or if a relatively high number of respondents will participate in a factorial survey, it might also be possible to cover the whole vignette universe (for an application of clustered random designs, see, e.g., Jasso and Rossi 1977:643; Jasso and Opp 1997:953; Beck and Opp 2001:293). The most extreme application of a clustered random design would be to restrict the survey to only one vignette sample. In this way, different sampling errors that otherwise would result from using more than one vignette sample will be leveled off by holding them constant (cf. also Jasso 2006:393). But a constant sampling error does not mean that there is no sampling error. If a random sample deviates strongly from the vignette universe and if a high number of respondents judged the vignettes over and above this, a generalization of the empirical results to the vignette universe might be biased by high errors, threatening the validity of the study. For this reason, drawing only one random vignette sample for all participants cannot be recommended. However, the view stated by Alves and Rossi (1978:544) that a generalization to the vignette universe would be lost without exception seems too pessimistic—at least for vignette samples very similar to the vignette universe.[4]

Quota Designs

Unlike simple random sampling, where the basic intent is to represent the vignette universe by different vignette samples, quota sampling tries to cover the vignette universe in central aspects by constructing only one vignette set. Whereas using only one random sample would leave the decision about which of the huge amount of possible vignette sets should be

used for all participants to chance, quota sampling borrows from the completely available knowledge about the statistical properties of the vignette universe for selecting the most suitable vignette sample.

In general, two different variants of *quota designs* can be distinguished: the more classical *fractional factorial designs* and *D-efficient designs*. In the following, the basic ideas for both types are outlined, starting with the more classical variant.

Fractional factorial designs. A central property of the vignette universe is that each possible level combination of different variables appears exactly one time. This characteristic does guarantee not only that all levels of a variable occur equally frequently (symmetrical or balanced) but also that the variables of different characteristics and all their interaction terms are mutually uncorrelated (orthogonal; cf. also Addelman 1962:23). In contrast to all other single samples, one sometimes will find a *fractional factorial design* that fulfills both criteria within the limits of an acceptable maximum number of vignettes per respondent (*symmetrical orthogonal design*). If no such design exists, one also might take into consideration using a fractional factorial design where the levels of each variable do not appear with equal frequency (asymmetrical). A necessary and sufficient condition for retaining the feature of mutual uncorrelatedness in such cases is that the levels of one variable occur with each level of the other variables with proportional frequency (*asymmetrical orthogonal design*; cf. Addelman 1962:23). All fractional factorial designs share the common property that at least the *b*-coefficients of all main effects of the vignette variables can be estimated as mutually uncorrelated for each individual respondent.

A reduction of the sample size is achieved for fractional factorial designs only by confounding (aliasing) main effects with higher order interaction effects (cf. Alexander and Becker 1978:96; Gunst and Mason 1991:48), assuming at the same time that the confounded interaction effects are negligible. A basic mathematical principle for constructing fractional factorial designs is the modular arithmetic applied to equation systems for the vignette variables (cf., for instance, Winer 1971:604-84; McLean and Anderson 1984). However, generating a fractional factorial design will be done in practice by using computer programs, such as SPSS (procedure "orthogonal design") or SAS (FACTEX procedure, %MktOrth macro, or %MktEx macro), or by consulting construction plans given by the literature (cf., for instance, Gunst and Mason 1991).[5] An advantage of consulting the literature above using computer programs such as SPSS is that it often includes an overview about whether only

main effects can be estimated uncorrelated (Resolution III designs); whether main effects can also be estimated uncorrelated with two-way interactions, whereby some two-way interactions are confounded with each other (Resolution IV designs); or whether main effects as well as two-way interactions can be estimated mutually uncorrelated (Resolution V designs; cf., for instance, Gunst and Mason 1991:48-9, 82-6; Kuhfeld 2005:50). This information might be important at least in situations where, for a given problem and a given sample size, designs with different resolutions are available. A higher resolution is at least preferable in cases where two-way interactions cannot be excluded on the basis of theoretical a priori knowledge.

D-efficient designs. Another variant of quota designs can be realized by somewhat relaxing the classical requirement of perfect orthogonality. The reason for shifting the main focus is stated by Kuhfeld et al. (1994) as follows: "Orthogonality is not the primary goal in design creation. It is a secondary goal, associated with the primary goal of minimizing the variance of the parameter estimates. Degree of orthogonality is an important consideration, but other factors should not be ignored" (p. 545). Since symmetrical orthogonal designs are balanced as well as orthogonal, they do not only represent the vignette universe most adequately but also minimize the variance of the parameter estimates. By choosing such designs as reference, *D-efficiency* has been proposed as a standard measure of goodness that captures both characteristics—balance and orthogonality—simultaneously. D-efficiency can be calculated according to the following formula (cf. also Kuhfeld et al. 1994:547):

$$\text{D-efficiency} = 100 \bullet \frac{1}{N_D \bullet \left|(X' \bullet X)^{-1}\right|^{\frac{1}{p}}} = 100 \bullet \left(\frac{1}{N_D} \bullet |X' \bullet X|^{\frac{1}{p}}\right), \quad (1)$$

where N_D denotes the set size of a design; $|X' \cdot X|$denotes the information matrix of the vignette variables, including the intercept; and p denotes the number of b-coefficients, including the intercept that have to be estimated. If all vignette variables are standardized orthogonally coded (cf. Kuhfeld 2005:65), then D-efficiency is scaled to range from 0 to 100 (cf. Kuhfeld et al. 1994:547, 549). Its maximum value of 100 will be reached only by symmetrical orthogonal designs. Hence, neither asymmetrical nor nonorthogonal designs will ever reach a D-efficiency of 100. Now, sometimes no symmetrical orthogonal exists. Whereas asymmetrical fractional factorial designs sacrifice in such situations perfect balance to preserve orthogonality,

search algorithms for D-efficient designs try to find an optimally efficient solution between perfect balance and orthogonality. For this reason, most D-efficient designs deviate at least slightly from orthogonality.

A D-efficiency value measures the goodness of a design relative to a symmetrical orthogonal design. Even if such a design may be far away from being possible for a given research question, the value 100 provides at least a rough reference for the goodness of a generated design. This applies at least to research questions where only qualitative variables will be analyzed. If quantitative variables that are not standardized orthogonally coded have to be included, then D-efficiency is no longer restricted to a maximum of 100, and the absolute value no longer has a clear interpretation. However, since calculating the relative D-efficiency measured by the ratio between the D-efficiency values of two competing designs does not require a special coding (although the same coding has to be used for both candidate designs), relative D-efficiency can be used in any case as a measure for the increase of efficiency due to preferring one design over another (cf. Kuhfeld et al. 1994:548-9). Since relative D-efficiency by itself is unaffected by the sample size, one can and should compare designs with different set sizes.

Finding a suitable D-efficient design requires search algorithms that can be provided only by computer programs such as JMP ("Custom Design"),[6] SAS (ADX "Optimal Design," the OPTEX procedure, or the %MktEx macro), or the conjoint value analysis (CVA) module of Sawtooth Software. Because nonexhaustive search algorithms are used, a computer program may fail to find *the* optimal design, even if the search algorithm is carried out several times (cf. Kuhfeld et al. 1994:547; Sawtooth Software 1997-2002:7-10). For this reason, the term *D-efficient design* is more appropriate than the also used term *D-optimal design*. A generated design's D-efficiency will be reported at least on demand by the three computer programs mentioned above.[7]

Arguments for or Against Random and Quota Sampling in Factorial Surveys

In the following, the main arguments for or against random and quota designs are discussed. Besides basic questions of *feasibility and applicability*, I address statistical issues of *efficiency and power* as well as methodological issues of *reliability and validity*.

Feasibility and Applicability

A main argument that can be raised against (simple) random designs refers to their feasibility (i.e., to the comparably higher *efforts and costs* that are required to produce the questionnaires). Drawing a unique vignette sample for each potential participant will usually be done by computer programs (see, e.g., Rossi 1979:179; Rossi and Anderson 1982:41; Jasso 2006:344) and, for that reason, should not be the problem. The same also applies to shuffling the vignettes before being presented to each respondent, a procedure preventing that possible order effects could cause systematic bias to the estimators. However, the task becomes somewhat more complicated by the requirement that an individual vignette set should neither include constants nor vignette variables that are linear combinations of other variables of the same vignette set. Especially small vignette sets, as well as variables with only few levels, are relatively often affected by such outcomes. Without correcting for such outcomes, one would lose the possibility of estimating respondent-specific *b*-coefficients for each individual vignette variable.

Whereas the feasibility argument against (simple) random designs has been relaxed by using computer-assisted interviews, it remains important for written questionnaires used for mail surveys. Generating the written vignettes and integrating the vignette sets into the questionnaires is much more time-consuming for (simple) random designs than for quota designs. Hence, using (simple) random designs is, in such cases, the more expensive choice. However, also by opting for a quota design, it is recommended to rely not only on one but also on several questionnaires, each one containing another permutation of the selected vignette set. In this way, the likelihood that the answer behavior might be systematically biased by possible order effects will be reduced. The same will be reached by shuffling the vignettes if the vignettes will not be integrated into the questionnaire but printed on small cards attached to the questionnaire. However, in this case, there is a chance that individual vignettes will be lost or will not be sent back. All in all, one will reduce the probability that order effects might occur at all by asking respondents to have a look at least at several vignettes before starting to answer the questions.

One central limitation on the use of fractional factorial designs is that *sometimes no design might exist* for a reasonable maximum number of vignettes per respondent. While the set size for random samples is, from the statistical point of view, only restricted by the number of *b*-coefficients to be estimated, fractional factorial designs are subject to much more

restrictive mathematical rules of divisibility applied to the number of levels per vignette variable. In general, it is easiest to find a suitable fractional factorial design in cases where all vignette variables have numbers of levels that are a power of 2 (cf. Gunst and Mason 1991:51) or at least possess an equal number of levels, including a power of it (cf. also McLean and Anderson 1984:26). Finding a feasible solution is much more difficult when each vignette variable has a different prime number of levels. In such situations, the required sample size often increases rapidly with each additional vignette variable. The same applies to research questions where the number of vignette variables and/or the number of levels per variable are relatively high.[8]

The applicability of quota designs has been relaxed by the availability of computer programs that allow generating D-efficient designs. A further option for enhancing the flexibility for constructing quota designs might sometimes also be to try different numbers of levels for some of the vignette variables. In many factorial surveys, the choice of the number of levels of a vignette variable is, within the limits given by theory, somewhat arbitrary. Hence, choosing four instead of three levels, or four instead of five levels (i.e., choosing a prime number or a power of it that is more in line with the number of the levels of the other vignette variables), might allow finding a suitable quota design within the limits of a maximal possible set size without degrading the goals of the experiment (cf. Gunst and Mason 1991:51-2).

A last point to be discussed here is the problem of the applicability of quota designs in cases where the vignette universe includes logically *impossible combinations* of vignette characteristics. Studies about topics such as justice of income (cf. Jasso and Rossi 1977:642; Alves and Rossi 1978:545) or the social status of families (cf. Nock 1982:104) often include education as well as occupation of fictitious vignette persons. Now, a minimum level of school education is required to be qualified for a certain occupation. If such restrictions exist, one is forced to exclude the respective combinations from the study. Removing impossible combinations from (simple) random designs generally only increases the correlation between affected variables, but by doing so, fractional factorial designs in general also lose orthogonality between variables that have not been affected as well as frequently between variables that have and have not been affected. For this reason, fractional factorial designs would not be applicable any longer. However, the greater flexibility of D-efficient designs might solve the problem. Although excluding impossible combinations will always reduce efficiency of a chosen design, the loss will

sometimes be very small for D-efficient designs (for an example, see Kuhfeld et al. 1994:551).

Statistical Issues: Efficiency and Power

Efficiency

The key for gaining insight into a design's *efficiency* is the statistical formula for estimating standard errors in multiple regression analyses (for trivariate regression, see Thome 1990:166-7; for the general case, see also Fox 1991:7-8):

$$\hat{\sigma}(b_1) = \sqrt{\frac{\sum_{i=1}^{n} e_i^2 \Big/ (n-k-1)}{\sum_{i=1}^{n} (X_{1i} - \bar{X}_1)^2 \bullet \left(1 - R^2_{X_1;X_2,X_3 \ldots X_k}\right)}} \qquad (2)$$

where

- $\hat{\sigma}(b_1)$ is the estimated standard error of the unstandardized regression coefficient of X_1;
- $\sum_{i=1}^{n} e_i^2/(n-k-1)$ is the estimated error variance $\hat{\sigma}_\varepsilon^2$—that is, the observed error variance divided by the number of the remaining degrees of freedom (n refers to the set size, k to the number of estimated b-coefficients);
- $\sum_{i=1}^{n} (X_{1i} - \bar{X}_1)^2$ is the variation of X_1 across vignettes 1 to n; and
- $R^2_{X_1;X_2,X_3 \ldots X_k}$ is the coefficient of multiple determination of the explanatory variables X_2 to X_k on X_1.

Formula (2) includes *three components* that are affected by the choice between different designs: the variation of the vignette variables, the coefficient of multiple determination among the vignette variables, and—at least to a certain degree—the estimated error variance.

The higher the *variation of the vignette variables* and the lower the *coefficient of multiple determination among the predictor variables*, the lower, ceteris paribus, will be the estimated standard error of a respective unstandardized regression coefficient. If each respondent has to judge a unique random sample, then the variation of the vignette variables, as well as the coefficient of multiple determination among the predictor variables, will vary from respondent to respondent. Hence, the b-coefficient of a

respective vignette variable would be estimated for each respondent with a different standard error. As a consequence, a part of the observed variation of the regression coefficients across respondents would be nothing else than essentially not explainable sampling variation of the vignette samples (cf. also Hox, Kreft, and Hermkens 1991:501). Standardizing the vignette samples has the advantage that it eliminates this source of error. This point is especially important for small set sizes, where the regression coefficients are typically estimated with large errors.

Using a standardized vignette sample for all respondents avoids not only the problem of different standard errors but requires at the same time also choosing among all possible designs the most efficient one (i.e., a design that allows estimating *b*-coefficients with the lowest standard errors). On the background of these considerations, it becomes already clear that quota samples are, for a given set size, more efficient than an average random sample.

Since fractional factorial designs ensure orthogonality, they reduce the coefficient of multiple determination among vignette variables to zero, whereby the right-hand term of the denominator of formula (2) reaches its maximum of 1. Search algorithms for D-efficient designs, on the other hand, try to optimize both components of the denominator simultaneously by relaxing the requirement for perfect orthogonality. The variation of a quantitative vignette variable, however, can only reach its maximum when its values are restricted to both extremes, whereby the ratio between both extremes would have to be quoted 1 to 1. Including only the lowest and the highest plausible values would be justified from a pure statistical point of view by the argument that two points are sufficient for estimating a linear relationship (cf. Kuhfeld et al. 1994:549). More levels are, from this point of view, only needed if the functional relationship is, for theoretical reasons, assumed to be nonlinear or if the relationship is at least unknown. However, it might be recommended sometimes for reasons of enhancing the similarity between the vignette world and the real world to include more than two levels of a quantitative variable into a vignette sample.

The third component of formula (2) is the *estimated error variance* $\hat{\sigma}_{\varepsilon}^2$, a term that depends at least partly on the set size and the number of predictor variables to be included in each separate regression equation. Choosing a high number of vignette variables for a small set size will leave relatively few degrees of freedom for estimating the error variance. The higher the number of the remaining degrees of freedom (i.e., the more vignettes are judged by each respondent), the lower the standard error of a

respondent-specific *b*-coefficient. An advantage of random designs, especially over fractional factorial designs, is that the set size can be chosen freely within the limits of the respondent's reasonableness. Thus, an average random sample of a higher set size may be statistically as efficient as or more efficient than a quota sample of a lower set size.

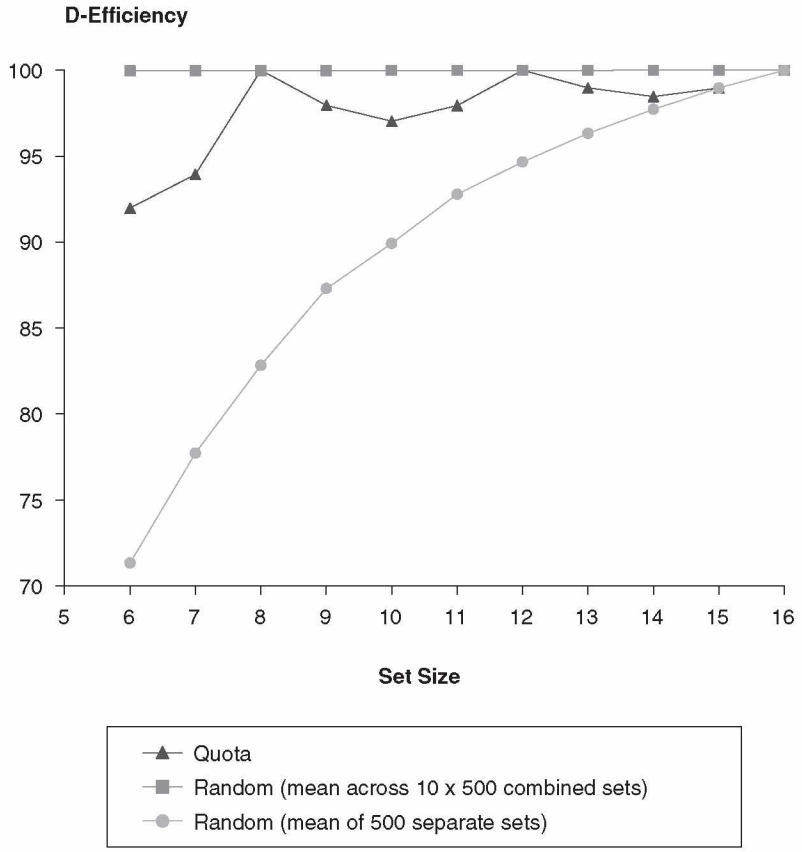
Our comparisons up to now did focus primarily on individual vignette sets and less on the combined vignette sample. Adding same quota samples will affect neither the degree of balance nor the mutual correlations among the vignette variables. For this reason, D-efficiency does not differ depending on whether it is computed for a single vignette set or across all vignettes included in a survey. Adding (simple) random designs, on the other hand, asymptotically reduces the mutual correlations among the vignette variables and increases the balance. Both characteristics affect D-efficiency. Although the mean D-efficiency of individual random samples will be comparably low, especially for relatively low set sizes, it will be much higher for the combined vignette sample. Which of both cases comes closer to reality depends on the heterogeneity of the respondents' answer behavior (i.e., on unexplained context effects).[9] Since unexplained context effects stemming from unmeasured respondent-level variables are very likely to occur in factorial surveys, random vignette designs can, as a rule, not fully benefit from their asymptotical characteristics.

The discussed relationship between set size and D-efficiency is illustrated in Figure 1 for a factorial survey consisting of four dichotomous vignette variables. Values for random designs are based on the more efficient method of drawing without replacement. The greater efficiency of drawing without replacement applies especially to relatively high set sizes, where the probability of drawing a vignette twice or more would otherwise be much higher than for lower set sizes. The upper line in Figure 1 depicts, for each set size, the mean D-efficiency across 10 times 500 combined sets (each of the 10 "surveys" without unexplained interrespondent heterogeneity); the lower one depicts, for each set size, the mean of 500 separate sets. Differences between random and quota designs will be more and more leveled off by increasing the set size.

Power

A design's D-efficiency is determined exclusively by the degree of orthogonality as well as the variation of the vignette variables. Hence, looking only at a design's D-efficiency does not take into account the

Figure 1
D-Efficiency of Random and Quota Designs Across Different Set Sizes (Four Dichotomous Vignette Variables)



impact of the set size as well as the number of respondents. Even if, for a given research question, a highly efficient design is available, the set size and/or the number of respondents might be too low for relatively weak effect sizes to become significant. In this case, lacking *power* might be responsible for disappointingly not rejecting H_0, although a respective

effect does exist in the population (thereby committing a Type II error, also denoted by β; cf. Cohen 1992:156).

The relationship between effect size, sample size, significance level α, and power $1-\beta$ is expressed by the following formula (cf. also Snijders and Bosker 1999:142):

$$\frac{\text{effect size}}{\text{standard error}} \approx (z_{1-\alpha} + z_{1-\beta}), \tag{3}$$

whereby $z_{1-\alpha}$ and $z_{1-\beta}$ are z scores associated with the indicated α and β values.

For decisions about the needed sample size, it is necessary to estimate the sample size that is required to achieve a specific power (say, $1-\beta=$ .80) for a given significance level (say, $\alpha=.05$)and a hypothesized effect size b (a priori power analysis; cf. also Hox 2002:177). For two-level hierarchical models, there are two kinds of sample sizes: the set size n and the sample size of the respondents N, with $N \cdot n$ being the total sample size of the vignettes. Carrying out a priori power analysis for multilevel analyses requires usually that quite a large number of parameters are assumed to be known (means, variances, and covariances of the predictor variables, as well as the variances and covariances of the random effects; cf. Bosker, Snijders, and Guldemond 2003:8). If this information can be gathered from earlier research or at least a reasonable guess can be made, one can use, for instance, PINT (cf. Bosker et al. 2003) for calculating the needed sample size for both levels of analysis. Otherwise, one should carry out a pilot study that may give an impression of the needed parameter values (cf. also Snijders and Bosker 1993:257).

Methodological Issues: Reliability and Validity

The last point to be addressed here is *reliability* and *validity*. Reliability depends on a design's efficiency—the higher it is, the lower the estimated standard error (i.e., the more precisely a regression coefficient can be estimated). Quota designs will be more reliable than random designs of the same set size if a relatively high number of unexplained context effects do exist—that is, if, besides the intercept, a comparably high number of slopes have to be estimated with their own random component (this condition does not allow a random design to profit so much from its asymptotical characteristics). The expected difference, under such circumstances, will be especially high for situations where for a relatively low set size, a

very high D-efficient quota design exists. If, on the other hand, relatively few or no unexplained context effects will be expected and only a comparably low D-efficient quota design exists for a given set size, then a random design will be the more reliable choice. The difference between both designs will, as already illustrated in Figure 1, all in all diminish with an increasing set size.

High reliability is only a necessary but not a sufficient condition for high validity. Even if a measurement turns out to be highly reliable, the empirical results might be biased and for that reason might be highly invalid. Although quota designs will reach higher reliability under the described circumstances, they are seen in general as a somewhat less valid choice than random designs. The reason for this view is the higher susceptibility of quota designs to biases caused by interaction effects between vignette variables that were not included. Fractional factorial designs are constructed by perfectly confounding (aliasing) main effects with higher order interaction effects. So, if nonnegligible unexpected interaction effects do exist and—due to the confounding pattern of the chosen design—cannot be estimated, not only the estimators of the constituting main effects of an interaction effect but also the estimators of the main effects with which the interaction is perfectly confounded will suffer from bias. Since higher order interaction effects will be found very rarely in social science (cf. Louviere 1988:40), such problems are in practice almost always restricted to two-way interactions between vignette variables. On the basis of these considerations, it should already be clear that if, for a given research question and a given set size, a higher resolution design exists, then it should be preferred over a lower resolution design. For D-efficient designs, the exact correlation structure between the terms for main and interaction effects is in general more complicated than for fractional factorial designs. Possible interaction terms have to be specified before the search algorithm starts.

An option to reduce the higher susceptibility of quota designs to possible biases might also be to increase the set size since, for higher set sizes, more interaction effects between vignette variables also can be estimated. Thus, increasing the set size will, all in all, reduce not only the differences between random and quota designs regarding their D-efficiency but also the differences with respect to potential biases. Another option that might help to prevent potential biases would be to use not only one, but, if they exist, several quota designs of the same efficiency that fulfill the requirements given by theory.

Example Mulitilevel Analyses on the Basis of Simulated Data

The following sample analyses are carried out to illustrate the differences between random and quota designs to be expected on the basis of our statistical and methodological considerations. For reasons of parsimony, we use again for our comparison a factorial survey consisting of four dichotomous variables. Since rather high unexplained interrespondent heterogeneity is very likely to be found in factorial surveys, our comparison will be based on a rather realistic situation where, besides the intercept, some slopes have to be estimated with their own random component. The intended comparison will be restricted to two different set sizes: At first, a relatively low set size will be chosen where a highly D-efficient quota design exists. This is given for a set size of 8 vignettes. Under these circumstances, the quota design should be more *efficient* than a random design of the same set size. Thus, the quota design should allow estimating the unstandardized regression coefficients with a lower standard error than the random design. As a consequence, t values computed as the ratio between b-coefficients and their standard error should, ceteris paribus, be higher for the quota design. Due to its higher *reliability*, the quota design should also be more suited for detecting unexplained interrespondent heterogeneity. However, the comparably lower *power* of the random design caused by its lower efficiency could at least be compensated by increasing the set size. But for such a situation, there will be also another quota design that might, despite its somewhat reduced D-efficiency, again turn out to be the more efficient choice. To see whether this assumption holds under the given circumstances, our comparison will be extended to a set size of 10 vignettes.

The advantages of quota designs are potentially endangered by their higher susceptibility to systematic biases that would reduce the *validity* of the results. In the following, it is assumed that vignette-level interaction effects do not exist. Besides b-coefficients, their standard errors, and t values, further measures that are central to multilevel analyses are also included in our statistical comparisons.

Data Basis

Producing Random and Quota Samples

The following comparison includes two quota and two random designs. Since the degree of D-efficiency depends on restrictions imposed by rules

of divisibility, one should in general try to find a set size where a high D-efficient quota design exists. In our case, a half-fractional factorial design (set size 8) of Resolution IV has been selected. Since the design is symmetrical and orthogonal, it reaches a D-efficiency of 100. Since no fractional factorial design exists for a set size of 10, a D-efficient design has been generated instead. The D-efficiency of the selected quota design produced without specifying interaction terms is 97.032. The coding plans for both selected quota designs are documented in Table 1. To get a realistic survey size, each quota design has been replicated 500 times. The needed random samples of set sizes 8 and 10 have been generated by drawing without replacement. In this way, a comparably higher D-efficiency will be reached for the individual vignette sets than one would have reached by drawing with replacement. The D-efficiency of the generated random designs, based on the mean of the separate sets, is 82.826 for the lower set size and 89.027 for the higher one. The values for the combined sets are 99.988 and 99.985, respectively.

In order to give an illustration for the chosen setting consisting of four dichotomous variables, let us assume that we were interested in Inglehart's (1990) value change theory. Materialistic and postmaterialistic value orientations have to be measured according to Inglehart by a ranking procedure. Applying this technique for measuring both value orientations has been criticized by authors such as Bürklin, Klein, and Ru ß (1996) for its lacking theoretical adequacy. According to the latter authors, there is no trade-off relationship between value preferences such as "protecting freedom of speech" and "fighting rising prices" that would justify a ranking procedure. An example vignette (see Table 2) may illustrate how both value orientations could be measured without the criticized restriction by asking respondents how much they would like to be governed by a party for which the listed goals are either not so important (code 0) or very important (code 1). By using a factorial survey, one also avoids the problem of response sets frequently observed for simple rating procedures, which have been criticized also for this reason by Inglehart (1997:116-7) as an unsuitable alternative for measuring both value orientations.[10]

Generating Data for the Simulated Answer Behavior

For factorial surveys, it is reasonable to expect unexplained context effects. Thus, we have to include into our regression equation, by which the fictitious respondents' answer behavior will be simulated, not only *b*-coefficients but also random terms by which unexplained context influences

Table 1
Coding Plans for the Selected Quota Designs

	Half-Fractional Factorial Design (Set Size 8)[a]				D-Efficient Design (Set Size 10)[b]			
Vignette Number	X_1	X_2	X_3	X_4	X_1	X_2	X_3	X_4
1	0	0	0	0	0	0	0	1
2	0	0	1	1	0	0	1	0
3	0	1	0	1	0	0	1	1
4	0	1	1	0	0	1	0	0
5	1	0	0	1	0	1	0	1
6	1	0	1	0	1	0	0	0
7	1	1	0	0	1	0	0	1
8	1	1	1	1	1	1	0	0
9	—	—	—	—	1	1	1	0
10	—	—	—	—	1	1	1	1

a. The means of X_1 to X_4 are 0.5, and the bivariate correlations among the variables are 0. The full factorial design can be divided into two half-fractional factorial designs of Resolution IV by the following equation system (for a general introduction into modular arithmetic, cf. also Winer 1971 or McLean and Anderson 1984):

$$x_1 + x_2 + x_3 + x_4 = 0, \text{mod} 2$$
$$x_1 + x_2 + x_3 + x_4 = 1, \text{mod} 2$$

The confounding pattern can be found easily by replacing code 0 by –1 (effect instead of dummy coding) and multiplying the X variables for computing the interaction terms (cf. also Jobson 1991:501).

b. The means of X_1, X_2, and X_4 are 0.5, and the mean of X_3 is 0.4. The correlation between X_1 and X_2 is 0.2, and the correlation between X_1 and X_4, as well as between X_2 and X_4, is –0.2. All other bivariate correlations are 0.

can be modeled. By computing the judgments, we know at the same time the respondents' true answers that will serve later on as a standard for our empirical comparisons. Since this contribution is concerned with comparing random to quota designs and not with examining theoretically expected cross-level interactions between respondent-level (Level 2) and vignette-level (Level 1) characteristics, no Level 2 characteristics have been included in our fictitious regression equation. The chosen multilevel regression equation is

$$Y_{ij} = (0 + u_{0j}) + 0.5 \bullet X_{1ij} + 1.5 \bullet X_{2ij} + (0.5 + u_{3j}) \bullet X_{3ij} + (1.5 + u_{4j}) \bullet X_{4ij} + r_{ij}, \quad (4)$$

Table 2
Vignette Example for Measuring Inglehart's Value Orientations

Political Goal:	*For the Governing Party the Goal:*
- Maintaining law and order in this nation	- not so important
- Giving people more say in government decisions	- not so important
- Fighting rising prices	- very important
- Protecting freedom of speech	- very important

I would like to be governed by such a party ...

1	2	3	4	5	6	7	8	9
not at all								*very strongly*

where i indicates a respective vignette and j a respective respondent. The intercept has a grand mean of 0. The grand mean values for the b-coefficients of the X variables are 0.5, 1.5, 0.5, and 1.5, respectively. Each fictitious respondent has her or his individual intercept as well as her or his individual b-coefficient for X_3 and X_4 that differ by the random components u_{0j}, u_{3j}, and u_{4j} from the respective grand coefficient. The error terms are distributed normally with a mean of 0. The standard deviations, also called *tau* for the respondent level, have been specified as 1.5, 0.8, and 1.2:

$$N(u_{0j}|\bar{u}_{0j}=0;\tau_0=1.5), N(u_{3j}|\bar{u}_{3j}=0;\tau_3=0.8), N(u_{4j}|\bar{u}_{4j}=0;\tau_4=1.2). \quad (5)$$

The error term r_{ij} for the vignette level is also normally distributed with a mean of 0 and a standard deviation of 1.5:

$$N(r_{ij}|\bar{r}_{ij}=0;\sigma_R=1.5). \quad (6)$$

To reduce the probability that the results of the simulation could be statistical outliers, not only 1 but 50 separate multilevel regressions were estimated for each design. For this purpose, both quota designs have simply been replicated. Since random designs differ from each other, it was necessary to generate the needed number of additional random designs. The needed number of additional variance components has been produced by using formulas (5) and (6). To minimize differences not genuine to a

respective design, the same 50 respondent-level error terms have been used for all four designs. Due to the different set size, two sets of vignette-level error terms have been generated for each of the 50 separate regression models. To enhance the comparability, the vignette-level error terms of the lower set size 8 have also been used for the higher set size 10. The lacking two error terms have been added to each vignette set, whereby means and standard deviations remain unchanged.

Empirical Results of the Example Analysis

All in all, 50 separate multilevel regressions have been estimated in HLM 6 for each of the four designs. On this basis, we calculated the means and standard deviations of the estimates that will be compared across the four designs. The results are presented in Table 3.

The estimated *b*-coefficients (first block; i.e., rows 1-5) are, on average for all four designs, very close to the expected values specified by formula (4). As long as no (unmodeled) vignette-level interaction effects are present, D-efficient designs also allow estimating regression coefficients without bias. The first systematic differences appear in the block for the estimated standard errors $\hat{\sigma}$ of the *b*-coefficients (second block). Due to its high *efficiency*, the fractional factorial design produces consistently at least slightly lower estimated standard errors than the random design of the same set size. However, the same does not apply to the comparison between the D-efficient design and the random design of set size 10. Here, both designs turn out to have, on average, nearly the same estimated standard errors for the *b*-coefficients. Thus, the random design of set size 10 can, under the given circumstances, already sufficiently profit from its asymptotical characteristics to catch up with the D-efficient design. As a consequence, the *reliability* of the *b*-coefficients of the same vignette variables is also nearly the same for both designs. The increased *power* of using a higher set size is reflected by their lower estimated standard errors. So, the random design of set size 8 produces the highest estimated standard errors, and both designs of set size 10 produce the lowest ones.

If the same *b*-coefficient is estimated with a lower standard error, then the *t* value will be larger, and the *b*-coefficient will become significant earlier. For this reason, the observed *t* values of a vignette variable's *b*-coefficients should follow very closely the reversed order of their estimated standard errors. Since both designs of set size 10 turned out to have nearly the same estimated standard errors for the *b*-coefficients of the

Table 3
Mean and Standard Deviation of Statistical Coefficients Across 50 Regressions, Each Estimated Separately for 500 Simulated Respondents

	Random Design (Set Size 8)		Quota Design (Set Size 8)		Random Design (Set Size 10)		Quota Design (Set Size 10)	
	Mean	Standard Deviation	Mean	Standard Deviation	Mean	Standard Deviation	Mean	Standard Deviation
$b_{Intercept}$	–0.004	0.046	–0.002	0.046	0.003	0.041	0.004	0.035
b_1	0.495	0.048	0.503	0.052	0.497	0.038	0.500	0.036
b_2	1.504	0.053	1.503	0.046	1.502	0.042	1.492	0.036
b_3	0.503	0.045	0.504	0.039	0.501	0.038	0.505	0.037
b_4	1.506	0.055	1.493	0.044	1.495	0.047	1.495	0.047
$\hat{\sigma}$ (Intercept)	0.086	0.005	0.085	0.002	0.083	0.001	0.083	0.001
$\hat{\sigma}(b_1)$	0.051	0.000	0.047	0.000	0.044	0.000	0.044	0.000
$\hat{\sigma}(b_2)$	0.051	0.001	0.047	0.000	0.044	0.000	0.044	0.000
$\hat{\sigma}(b_3)$	0.062	0.003	0.060	0.002	0.057	0.002	0.056	0.002
$\hat{\sigma}(b_4)$	0.073	0.002	0.072	0.002	0.069	0.002	0.069	0.001
$t_{Intercept}$	–0.045	0.527	–0.023	0.536	0.034	0.492	0.051	0.416
t_{b1}	9.782	0.960	10.656	1.122	11.224	0.863	11.387	0.820
t_{b2}	29.754	1.053	31.814	1.005	33.957	0.939	33.975	0.882
t_{b3}	8.175	0.817	8.479	0.693	8.851	0.765	9.027	0.664
t_{b4}	20.626	0.981	20.872	0.846	21.692	0.914	21.643	0.872

$\hat{\tau}_0$	1.500	0.050	1.502	0.043	1.505	0.035	1.500	0.035
$\hat{\tau}_3$	0.804	0.082	0.808	0.064	0.801	0.061	0.793	0.065
$\hat{\tau}_4$	1.199	0.066	1.201	0.059	1.193	0.057	1.193	0.043
$\hat{\sigma}_R$	1.497	0.014	1.494	0.011	1.499	0.010	1.500	0.009
χ^2_{U0}	1,760.783	99.978	1,845.118	80.915	2,125.523	86.168	2,370.624	97.032
χ^2_{U3}	748.240	55.126	792.174	46.850	825.901	51.786	836.052	55.951
χ^2_{U4}	1,046.433	66.730	1,145.644	69.367	1,221.742	72.596	1,288.437	60.270
R^2_1	.185	.008	.184	.008	.183	.008	.156	.006
R^2_2	.026	.013	.000	.000	.015	.011	.000	.000
Iterations	15.900	4.670	3.000	0.000	11.240	2.255	3.000	0.000

same vignette variables, they also have nearly the same t values for the respective b-coefficients (cf. third block). Lower t values are consistently observed for both designs of set size 8, whereby the random design produces the poorest results.

The mean values for the estimated standard deviations $\hat{\tau}$ and $\hat{\sigma}$ of the error terms u and r (cf. fourth block) are very close to the values specified in formulas (5) and (6). Larger differences between the designs become visible in the fifth block, where the results for the chi-square tests are documented. The chi-square test is used in HLM to test whether an estimated variance component $\hat{\tau}^2$ of an error term u becomes significant. A chi-square value is computed by summing up across all respondents the squared deviation of a respondent-specific computed b-coefficient from its overall estimate computed across all respondents divided by the respondent-specific estimated sampling variance of that b-coefficient (i.e., by the square of the respondent-specific estimated standard error of that b-coefficient; cf. also Hox 2002:43). As long as the estimated standard error for a b-coefficient is—ceteris paribus—higher for a random design than for a quota design, the respective chi-square value will be lower. Thus, for a given set size, such a random design would be less able to discover unexplained respondent heterogeneity. Increasing the set size reduces the estimated standard errors for the b-coefficients—a respective chi-square value becomes higher. On the background of these considerations, it becomes clear why the empirical results show the lowest chi-square values for the random design of set size 8 and the highest ones for both designs of set size 10.

The average coefficients of multiple determination computed according to the simplified formulas proposed by Snijders and Bosker (1994:350-54; cf. also Snijders and Bosker 1999:99-105) are documented immediately below the chi-square values. By including the four vignette variables, roughly 18.4 percent of the variance is explained at Level 1 for all except the D-efficient design. Thus, the reduced variation of one of the X variables and the nonzero correlations between some of the X variables are together responsible for obtaining a comparably somewhat lower R_1^2 of only 15.6 percent for the D-efficient design. Since no respondent-level variables have been included in the regression equation, the R^2 at Level 2 should be 0. This expectation is only fulfilled by both quota designs.[11] For the random designs, the explained Level 2 variance amounts to 2.6 and 1.5 percent, respectively. In contrast to quota designs, where each individual participant receives the same vignette set, at least slightly different vignette sets are used for random designs. Now, the differences between the means of the vignette variables of different vignette sets also cause

variation in the average of the dependent variable between respondents. Including the vignette variables into regression analysis explains these differences between respondents and, for that reason, increases the Level 2 coefficient of determination. Thus, the fractional factorial design, all in all, produces the best estimators for both coefficients of determination.

A final look at the average number of iterations illustrates that, due to its more complex error structure caused by using a unique vignette set for each respondent, random designs need many more iterations until the maximum likelihood function for estimating multilevel regressions converges. Nonetheless, the number of iterations is always well below 100, the default value given by HLM that should not be exceeded too much.

Conclusion

In contrast to conjoint analyses that are carried out almost exclusively on the basis of quota designs, it is common to use random designs for vignette analyses. The pros and cons for preferring random designs over quota designs are usually not discussed at all. Most of the time, not even a hint can be found that quota designs could be an attractive alternative to random designs. The purpose of this article was to contribute filling this gap.

From the point of feasibility, costs and efforts are seen as the central arguments against using random designs in representative survey research. The situation, however, has been relaxed at least for computer-assisted interviews. Whereas random samples can be generated for each desired set size, fractional factorial designs are frequently not available within the limits of a reasonable set size. Here, the situation has been relaxed since computer programs for generating D-efficient designs are available. From a statistical point of view, higher efficiency, higher reliability, and higher power are seen as main arguments for favoring quota designs over random designs. These arguments, however, apply mainly to situations where, for a relatively low set size, a highly D-efficient design is available and where, at the same time, relatively high unexplained interrespondent heterogeneity is expected.

Quota designs are seen generally as less valid than random designs. This argument, however, probably only applies to low-resolution fractional factorial designs, where main effects are already confounded with two-way interaction effects, as well as to D-efficient designs, where a higher efficient design has been chosen over a lower efficient design that would have allowed estimating relevant interactions. From this point of view, one should think seriously about possible interaction effects. To

reduce the higher susceptibility of quota designs to systematic bias, one should use at least several permutations of a selected design. For the same reason, it might be recommended that one should also use several quota designs of the same D-efficiency. Since vignette analysis is a complex measurement, it is recommended in any case to carry out a pretest before starting the main survey. If these considerations are taken seriously, then quota designs might become an attractive alternative to random designs. This, however, applies—as already stated—probably above all to situations where relatively small vignette sets have to be used and where, besides the intercept, a rather high number of slopes have to be estimated with their own random component.

Notes

1. The differences between vignette and conjoint analysis can be put down primarily to their origin as well as to the commonly used statistical methods for analyzing the data: While in social sciences, predominantly the term *vignette analysis* prevails, in economic sciences—particularly in marketing research—the term *conjoint analysis* dominates. Since marketing is primarily interested in the preference order for certain products, the dependent variable is mostly measured by a ranking task. Hence, the variable reaches ordinal scale level. Because marketing research is mainly interested in market choice behavior and market segmentation, standardized part-worth utilities computed on the basis of the unstandardized regression coefficients from the individual respondents are frequently used for carrying out cluster analyses (although ordinary least squares [OLS] regression is inappropriate for rank-order data, it has been consistently found in a number of studies that nonmetric estimation does not appear to give substantially better results than metric procedures; cf. Vriens 1995:64-5). For research questions in social science, strict rank orders are mostly unnecessary. Hence, the dependent variable is mostly measured by using a rating task. For that reason, the variable is assumed to reach metric scale level. To test hypotheses about the relationship between predictors (vignette as well as respondent characteristics) and the dependent variable, researchers choose regression analysis most often in vignette analysis. Except for differences in terminology, in setting their main focus of interest, and in selecting their mainly used procedures for analyzing data, in principle, there are probably no differences between vignette and conjoint analyses.

2. For conjoint analyses, a short discussion can be found in Green and Srinivasan (1978:109-11).

3. The main focus of this article is on research questions to assess the impact of vignette and respondent characteristics on the respondent's judgment behavior. Although such relationships are by far the most analyzed ones in the social sciences, it needs to be mentioned that factorial surveys also allow exploring the consequences of the respondent-specific estimates of the vignette variables (cf. also Jasso 2006). However, since no simultaneous estimation programs such as HLM are available for such cases, one would have to carry out in a first step a separate OLS regression for each individual respondent and use the resulting *b*-coefficients thereafter in a separate second step as predictors for a dependent respondent-level variable.

4. The exact wording of Alves and Rossi (1978) is as follows: "While there might have been much to gain from standardizing the sample of vignettes, presenting to each respondent the same set, at the same time we would have lost the ability to generalize to the universe of all possible vignettes" (p. 544).

5. Further references for constructing balanced orthogonal fractional factorial designs can be found at http://support.sas.com/techsup/technote/ts723.html (2.12.2005).

6. A restriction of JMP 5.1 is that it does not allow including quantitative variables with more than two levels into a vignette set.

7. Although D-efficiency is the most usual measure for a design's efficiency, it is not the only one. Kuhfeld et al. (1994:546-7), for instance, also refer to A- and G-efficiency. All three of these criteria are convex functions of the eigenvalues of $(X' \cdot X)^{-1}$ and hence are usually highly correlated.

8. A reasonable set size is, according to Jasso (2006:343), in the range of 40 to 60 vignettes. However, the recommended set size might also depend on respondent characteristics: Respondents of a representative national population sample may become fatigued or bored much earlier than respondents of a highly motivated student sample. Another point is that the costs for a survey will rise with an increasing set size. Hence, Beck and Opp (2001:291) recommend for most representative surveys a set size in the range of 10 to 20 vignettes. The number of vignette variables should in general not exceed a maximum of six variables, each consisting of up to four or five levels (these upper limits can at least be found for conjoint analysis; cf. Green and Srinivasan 1978:108). If fewer vignette variables are included, one might also use more levels and vice visa. Finally, to get reliable estimates for the individual respondents, one should choose a set size at least 1.5 times higher than the number of parameters to be estimated (cf. Sawtooth Software 1997-2000:7).

9. More precisely: depending on existing interrespondent heterogeneity that has to be modeled in multilevel analyses by additional random components, allowing respondents to have their own intercept and/or their own *b*-coefficients that deviate "randomly" from the estimate of the respective grand mean. "Randomness" in multilevel analyses can be regarded as representing the effects of unmeasured Level 2 variables (respondent level) and hence may be interpreted as unexplained Level 2 variability (interrespondent heterogeneity; cf. also Snijders and Bosker 1999:45).

10. Empirically, however, it can be shown that Inglehart's (1997) simple ranking procedure is *no* less susceptible to response sets than the simple rating procedure (Klein et al. 2004).

11. The mean of R_2^2 is in both cases slightly negative (the exact values are –.00001 for the fractional factorial design and –.000005 for the D-efficient design). Because negative coefficients of determination do not make sense, the negative sign has been dropped in Table 3.

References

Addelman, Sidney. 1962. "Orthogonal Main-Effect Plans for Asymmetrical Factorial Experiments." *Technometrics* 4 (1): 21-46.

Alexander, Cheryl S. and Henry J. Becker. 1978. "The Use of Vignettes in Survey Research." *Public Opinion Quarterly* 42 (1): 93-104.

Alves, Wayne M. and Peter H. Rossi. 1978. "Who Should Get What? Fairness Judgments of the Distribution of Earnings." *American Journal of Sociology* 84 (3): 541-64.

Beck, Michael and Karl-Dieter Opp. 2001. "Der faktorielle Survey und die Messung von Normen." *Kölner Zeitschrift für Soziologie und Sozialpsychologie* 53 (2): 283-306.

Bosker, Roel J., Tom A. B. Snijders, and Henk Guldemond. 2003. *PINT: Estimating Standard Errors of Regression Coefficients in Hierarchical Linear Models for Power Calculations* [User's manual]. Retrieved December 2, 2005, from http://stat.gamma.rug.nl/snijders/

Bürklin, Wilhelm, Markus Klein, and Achim Ru ß. 1996. "Postmaterieller oder anthropozentrischer Wertewandel? Eine Erwiderung auf Ronald Inglehart und Hans-Dieter Klingemann." *Politische Vierteljahresschrift* 37 (3): 517-36.

Cohen, Jacob. 1992. "A Power Primer." *Psychological Bulletin* 112 (1): 155-9.

Fox, John. 1991. *Regression Diagnostics: An Introduction*. Sage University Paper Series on Quantitative Applications in the Social Sciences, 07-079. Newbury Park, CA: Sage.

Green, Paul E. and V. Srinivasan. 1978. "Conjoint Analysis in Consumer Research: Issues and Outlook." *Journal of Consumer Research* 5:103-23.

Gunst, Richard F. and Robert L. Mason. 1991. *How to Construct Fractional Factorial Experiments: Vol. 14. The ASQC Basic References in Quality Control: Statistical Techniques*. Milwaukee, WI: ASQC Quality Press.

Gustafsson, Anders, Andreas Herrmann, and Frank Huber. 2000. "Conjoint Analysis as an Instrument of Market Research Practice." Pp. 5-45 in *Conjoint Measurement: Methods and Applications*, edited by A. Gustafsson, A. Herrmann, and F. Huber. Berlin: Springer.

Hox, Joop. 2002. *Multilevel Analysis: Techniques and Applications*. Mahwah, NJ: Lawrence Erlbaum.

Hox, Joop J., Ita G. G. Kreft, and Piet L. J. Hermkens. 1991. "The Analysis of Factorial Surveys." *Sociological Methods & Research* 19 (4): 493-510.

Inglehart, Ronald. 1990. *Cultural Shift in Advanced Industrial Society*. Princeton, NJ: Princeton University Press.

———. 1997. *Modernization and Postmodernization: Cultural, Economic, and Political Change in 43 Societies*. Princeton, NJ: Princeton University Press.

Jasso, Guillermina. 1988. "Whom Shall We Welcome? Elite Judgments of the Criteria for the Selection of Immigrants." *American Sociological Review* 53:919-32.

———. 2006. "Factorial Survey Methods for Studying Beliefs and Judgments." *Sociological Methods & Research* 34 (3): 334-423.

Jasso, Guillermina and Karl-Dieter Opp. 1997. "Probing the Character of Norms: A Factorial Survey Analysis of the Norms of Political Action." *American Sociological Review* 62:947-64.

Jasso, Guillermina and Peter H. Rossi. 1977. "Distributive Justice and Earned Income." *American Sociological Review* 42:639-51.

Jobson, J. D. 1991. *Applied Multivariate Data Analysis: Vol. I. Regression and Experimental Design*. New York: Springer.

Klein, Markus, Hermann Dülmer, Dieter Ohr, Markus Quandt, and Ulrich Rosar. 2004. "Response Sets in the Measurement of Values: A Comparison of Rating and Ranking Procedures." *International Journal of Public Opinion Research* 16 (4): 474-83.

Kuhfeld, Warren F. 2005. "Experimental Design, Efficiency, Coding, and Choice Designs." Pp. 47-97 in *Marketing Research Methods in SAS: Experimental Design, Choice, Conjoint, and Graphical Techniques*, edited by W. F. Kuhfeld. Retrieved December 2, 2005, from http://support.sas.com/techsup/tnote/tnote_stat.html

Kuhfeld, Warren F., Randall D. Tobias, and Mark Garratt. 1994. "Efficient Experimental Design With Marketing Research Applications." *Journal of Marketing Research* 31:

545-57. (A revised version of the paper can be downloaded from http://support.sas.com/techsup/tnote/tnote _stat.html)

Louviere, Jordan J. 1988. *Analyzing Decision Making: Metric Conjoint Analysis*. Sage University Paper Series on Quantitative Applications in Social Sciences, 07-067. Newbury Park, CA: Sage.

Marshall, Pablo and Eric T. Bradlow. 2002. "A Unified Approach to Conjoint Analysis Models." *Journal of the American Statistical Association* 97 (459): 674-82.

McLean, Robert A. and Virgil L. Anderson. 1984. *Applied Factorial and Fractional Designs*. New York: Marcel Dekker.

Nock, Steven L. 1982. "Family Social Status: Consensus on Characteristics." Pp. 95-118 in *Measuring Social Judgments: The Factorial Survey Approach*, edited by P. H. Rossi and S. L. Nock. Beverly Hills, CA: Sage.

Rossi, Peter H. 1979. "Vignette Analysis: Uncovering the Normative Structure of Complex Judgments." Pp. 176-86 in *Qualitative and Quantitative Social Research: Papers in Honor of Paul F. Lazarsfeld*, edited by R. K. Merton, J. S. Coleman, and P. H. Rossi. New York: Free Press.

Rossi, Peter H. and Andy B. Anderson. 1982. "The Factorial Survey Approach: An Introduction." Pp. 15-67 in *Measuring Social Judgments: The Factorial Survey Approach*, edited by P. H. Rossi and S. L. Nock. Beverly Hills, CA: Sage.

Rossi, Peter H. and Steven L. Nock. 1982. "Preface." Pp. 9-13 in *Measuring Social Judgments: The Factorial Survey Approach*, edited by P. H. Rossi and S. L. Nock. Beverly Hills, CA: Sage.

Sawtooth Software. 1997-2002. CVA: A Full-Profile Conjoint Analysis System. Version 3. Sequim: Sawtooth Software. Retrieved December 2, 2005, from http://www.sawtooth-software.com/download/techpap/cva3tech.pdf

Snijders, Tom A. B. and Roel J. Bosker. 1993. "Standard Errors and Sample Sizes for Two-Level Research." *Journal of Educational Statistics* 18 (3): 237-59.

———. 1994. "Modeled Variance in Two-Level Models." *Sociological Methods & Research* 22 (3): 342-63.

———. 1999. *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling*. London: Sage.

Thome, Helmut. 1990. "Grundkurs Statistik für Historiker. Teil II: Induktive Statistik und Regressionsanalyse." *Historische Sozialforschung*, Supplement No. 3. Cologne, Germany: University of Cologne.

Vriens, Marco. 1995. *Conjoint Analysis in Marketing: Developments in Stimulus Representation and Segmentation Methods*. Capelle a/d IJssel: Labyrint Publication.

Winer, B. J. 1971. *Statistical Principles in Experimental Design*. 2nd ed. New York: McGraw-Hill.

Hermann Dülmer is an assistant professor of sociology at the Central Archive for Empirical Social Research of the University of Cologne. His content-related research is currently focused on comparative value research, including value change; electoral research, with a particular emphasis on right-wing extremism; and ethnocentrism and its determinants. His methodological interests focus on factorial surveys and on multilevel analysis. A recent journal article he wrote in collaboration with Markus Klein appeared in the *European Journal of Political Research* (2005). Together with colleagues, he wrote a monograph on the German national election in 2002 (Güllner, Dülmer, Klein, Ohr, Quandt, Rosar, and Klingemann 2005).